Transformer-Based Multi-Sensor Data Fusion for Airborne Surveillance Systems Using Adaptive Covariance Intersection

Gaurav Sharma[1,2*] and CRS Kumar[1,†]
R Rajesh[2,‡]

1 Defence Institute of Advanced Technology (DIAT), Girinagar, Pune, Maharashtra 411025, India
2 Centre for Airborne Systems (CABS), Defence Research and Development Organisation, Wind Tunnel Road, Bengaluru, Karnataka 560064, India

Abstract

Multi-sensor data fusion involving heterogeneous sensors is a challenging task. The problem statement becomes more complex in case of airborne surveillance as continuous platform motion, aggressive maneuvering, and dense operational scenarios having multiple closely spaced, crossing, and coordinated targets, increase the likelihood of ambiguous data association and track fusion. This paper presents an end-to-end framework that fuses radar and identification-friend-or-foe observations for airborne surveillance systems through a transformer-based association pipeline coupled with an adaptive covariance-intersection fusion stage and a rule-based track lifecycle manager. The association stage learns spatial and temporal correspondences between asynchronous sensor scans using attention mechanisms. In contrast, the fusion stage learns per-scan trust weights rather than relying on constant mixing coefficients. Evaluated on a realistic, densely populated held-out scenario and benchmarked against classical nearest-neighbour and fixed-weight fusion baselines, the proposed framework demonstrates markedly improved detection reliability and comparable false-track behaviour, while surfacing open challenges in covariance calibration and track fragmentation that motivate continued refinement of the framework.

---

*Doctoral Research Scholar, Department of Computer Science and Engineering, Defence Institute of Advanced Technology, Girinagar, Pune. Maharashtra, India
† Professor, Department of Computer Science and Engineering , Defence Institute of Advanced Technology, Girinagar, Pune. Maharashtra, India
‡ Scientist G , Centre for Airborne Systems, DRDO, Wind Tunnel Road, Bengaluru, India

Nomenclature

L_{tr} = spatial (track–radar) association logit matrix, dimensionless
L_{ri} = temporal (radar–IFF) association logit matrix, dimensionless
N_t, N_r, N_i = number of predicted tracks, radar detections, and IFF detections, count
P = state error-covariance matrix, m² (position terms), (m/s)² (velocity terms)
P_f = fused error-covariance matrix, m²
x_{pred}, x_r, x_i, x_f = predicted, radar, IFF, and fused state vectors, m (position), m/s (velocity)
ω_{pred}, ω_r, ω_i = learned trust weights for prediction, radar, and IFF, dimensionless
A_{min} = minimum combined hit-score for track confirmation, dimensionless
M_{max} = maximum permissible consecutive missed scans before disassociation, scans

σ_ρ = range-measurement noise standard deviation, m
σ_θ, σ_φ = elevation and azimuth angular noise standard deviation, deg
σ_v = velocity-measurement noise standard deviation, m/s
$\chi 2$ = chi-squared statistic used for gating, dimensionless
ω = fixed mixing weight of classical covariance intersection, dimensionless

I. INTRODUCTION

Aerial target tracking using radar and Identification Friend-or-Foe (IFF) sensors is a core capability of airborne surveillance systems [16]. Radar produces six-dimensional kinematic estimates of position and velocity, while IFF contributes five-dimensional measurements combining position, speed, and an identity code that is not available from radar alone. The two sensors operate asynchronously, with scan intervals that do not align, which creates a structurally different fusion problem from synchronous multi-radar architectures [14]. Incommensurable measurement dimensionalities and unknown inter-sensor correlations further complicate the problem, motivating fusion frameworks that can reason jointly over heterogeneous, intermittently available observations rather than treating each sensor stream in isolation.

Classical probabilistic and multiple-hypothesis association methods [1, 2] remain widely used but degrade in dense, cluttered environments where the number of plausible track-to-detection pairings grows combinatorially. Fixed-model Kalman filters [4, 6] are computationally efficient but violate the stationarity assumptions on which they rely whenever a target maneuvers aggressively, since a constant-velocity or constant-turn model cannot simultaneously represent straight flight and evasive maneuvering. Covariance Intersection (CI) [7] offers a principled way to fuse estimates with unknown correlation, guaranteeing a consistent fused covariance without requiring knowledge of the cross-covariance between sources; however, its conventional formulation uses a single fixed mixing weight, which is suboptimal when the relative reliability of the contributing sensors varies over time and across the operating envelope.

Learning-based approaches have started to address these limitations. Verma et al. [9] demonstrated in his work that a Bi-LSTM sequence model can outperform a JPDA-IMM baseline by a substantial margin on single-radar tracking, but the work did not address the asynchronous, cooperative-sensor fusion problem central to this work. TrackFormer [11] and related transformer-based multi-object trackers [3, 12] have shown that attention mechanisms are well suited to data association, and TMSHF [14] extends this idea to multi-sensor hybrid fusion, but these frameworks presume synchronous sensor scans and do not model the differential scan-rate, differential-reliability

structure of a radar-IFF pair. Our prior work [17] introduced a two-stage transformer-based association module comprising a spatial Association Module and a cross-modal Temporal Association Module, trained with ground-truth-supervised pre-training followed by self-supervised fine-tuning, and demonstrated substantial gains in association accuracy on a moderate-density scenario. However, [17] focused exclusively on the association stage: the fusion weights were held fixed at their classical values, and principled track lifecycle management (birth, confirmation, disassociation, and deletion under asynchronous sensor gaps) was identified as future work rather than addressed.

The present paper closes this gap by extending [17] to a complete, end-to-end multi-sensor data fusion framework that couples the transformer-based association stage with a learned, adaptive Covariance Intersection fusion module and a principled, sensor-aware track lifecycle manager. The complete pipeline is evaluated on a substantially denser and more challenging scenario than considered previously. The primary contributions of this work are as follows:

1) A spatial association module implemented as a sparse Mahalanobis-gated spatial transformer with adaptive chi-squared thresholds, trained with a multi-term ground-truth-supervised loss.

2) A temporal association module using history-conditioned cross-modal attention with a persistent association-memory buffer and an identity-temporal regularization term.

3) A fusion and state-estimation module that replaces the fixed mixing weight of classical Covariance Intersection with a multilayer-perceptron trust head producing learned, context-conditioned per-scan weights, trained through a staged pre-training and joint fine-tuning curriculum.

4) A track manager implementing sensor-aware lifecycle management, including hit-score-based confirmation and differential per-sensor disassociation, that enables reliable track maintenance under asynchronous and intermittently missing sensor data.

The remainder of the paper is organized as follows. Section II describes the five-stage architecture of the fusion framework. Section III presents the neural network design and the three-phase training methodology used to obtain the reported model. Section IV presents the experimental evaluation, including the simulation scenario, the sensor noise model, comparison against classical baselines, and a discussion of filter consistency and computational performance. Section V concludes the paper and outlines directions for future work.

II. MSDF FRAMEWORK ARCHITECTURE

The proposed framework integrates asynchronous radar and IFF sensor tracks into unified system tracks through five functional stages operating in a closed feedback loop (Fig. 1). The pipeline executes at each radar scan event. Within each step, the temporally nearest IFF scan is retrieved through nearest-neighbor time-matching. The spatial association module fires on every radar step, while the temporal cross-modal association module fires conditionally, when a fresh IFF scan has occurred since the previous pipeline execution. This design means that IFF data is ingested within a radar-scan-triggered pipeline pass, so that the pipeline faithfully follows the asynchronous scan schedule of the two sensors rather than forcing them onto a common synchronous grid, which would otherwise require either discarding information or introducing artificial latency.

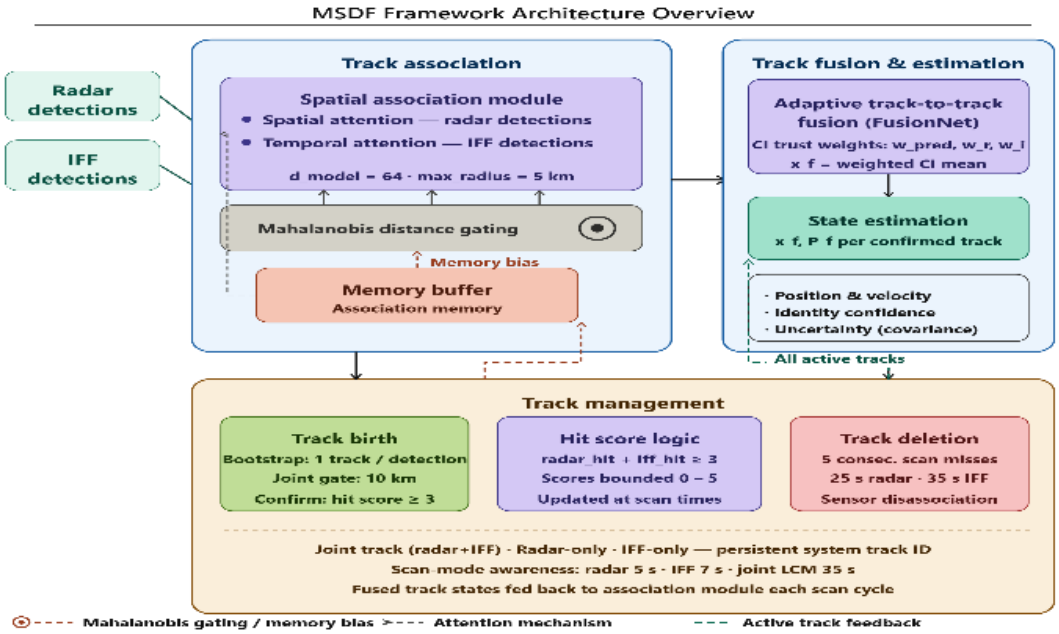
MSDF Framework Architecture Overview
Radar detections
IFF detections
Track association
Spatial association module
Spatial attention — radar detections
Temporal attention — IFF detections
d_model = 64 · max radius = 5 km
Mahalanobis distance gating
Memory bias
Memory buffer
Association memory
Track fusion & estimation
Adaptive track-to-track fusion (FusionNet)
CI trust weights: w_prad, w_r, w_i
x_f – weighted CI mean
State estimation
x_f, P_f per confirmed track
· Position & velocity
· Identity confidence
· Uncertainty (covariance)
All active tracks
Track management
Track birth
Bootstrap: 1 track / detection
Joint gate: 10 km
Confirm: hit score ≥ 3
Hit score logic
radar_hit + iff_hit ≥ 3
Scores bounded 0 – 5
Updated at scan times
Track deletion
5 consec. scan misses
25 s radar · 35 s IFF
Sensor disassociation
Joint track (radar+IFF) · Radar-only · IFF-only — persistent system track ID
Scan-mode awareness: radar 5 s · IFF 7 s · joint LCM 35 s
Fused track states fed back to association module each scan cycle
Mahalanobis gating / memory bias
Attention mechanism
Active track feedback

Fig. 1 Five-stage closed-feedback-loop architecture of the proposed multi-sensor data fusion framework.

A. Kinematic Prediction

Before association, each system track is propagated forward using a constant-velocity state-transition model. Process noise is intentionally omitted at this stage and instead added within the track manager after measurement association, thereby preventing unnecessary covariance inflation during coasting periods when a track is temporarily unobserved by one or both sensors. A fast-moving target can travel in the order of a kilometre between consecutive radar scans. Without instituting good prediction module , a gating test on raw, unpropagated track states would reject the correct detection purely because of platform motion, not because of an association ambiguity. Kinematic prediction

is therefore a necessary precursor to association rather than an optional refinement, and its accuracy directly bounds the tightness of the gate that the association stage can subsequently apply.

B. Spatial Association

The spatial association module maps the predicted track states and the radar detections, each represented with a six-dimensional mean and covariance, to a logit matrix that is resolved with the Hungarian assignment algorithm. An additional row and column represent learnable null tokens, enabling graceful handling of missed detections without forcing every track or detection into an assignment. A Mahalanobis-gated sparse candidate generator pre-filters the dense track-detection matrix to a small candidate set per track. With the chi-squared gate scaled according to each track's current positional uncertainty, together with a fallback rule non-empty candidate set are guaranteed even for tracks whose uncertainty briefly exceeds the nominal gate. Pointer-head bias regularization and a null-score anchor term discourage the network from collapsing onto an all-null solution, a degenerate but easily-reached local minimum in which the network trivially avoids every assignment decision. Multi-head self-attention then reasons globally across all track-detection pairs simultaneously, which allows the module to resolve competing claims on the same detection in a way that a purely local, pairwise gating rule cannot.

C. Temporal Association Module

This module resolves cross-modal radar-IFF matching across the asynchronous IFF scan boundary. It reads a padded track history in which each entry contains the kinematic state together with the last observed IFF identity code along with associated confidence value. Temporal attention allows the transformer to recognize that a track consistently reporting a particular identity code as the likely match for an incoming detection carrying the same code, even when positional ambiguity alone would be insufficient to disambiguate the pairing. A persistent association-memory buffer retains per-track identity-assignment priors across several scans, while suppressing oscillation in the assigned identity during periods when IFF is temporarily unavailable. An identity-temporal regularization term prevents the corresponding attention weights from collapsing during radar-only timesteps, when IFF provides no training gradient. Without this term the temporal pathway would otherwise be free to drift arbitrarily whenever it is not actively being supervised.

D. Global Joint Assignment

The spatial and temporal logit matrices both enter a global joint-assignment step that applies the Hungarian algorithm to produce a single, globally consistent one-to-one assignment: each system track receives at most one radar

detection and at most one IFF measurement per scan, and each detection is assigned to at most one track. This joint formulation is the key distinction from sequential or greedy alternatives, in which radar and IFF association decisions would be made independently and then reconciled heuristically. Under the joint formulation, a track that is only weakly supported by radar but strongly confirmed by its identity code can still receive a correct assignment, because both signals contribute evidence to the same optimization rather than being adjudicated in isolation. Unmatched radar detections trigger tentative track births, and miss counters for all unmatched tracks are incremented by the track manager.

E. Adaptive Covariance-Intersection Fusion

The fusion module combines three information sources for each confirmed joint track , the kinematically predicted state, the associated radar measurement, and the associated IFF measurement, each carrying its own covariance matrix. Rather than using the fixed mixing weight of classical Covariance Intersection, a trust-head multilayer perceptron maps a feature vector encoding the predicted state, both sensor measurements, the covariance log-diagonals, the velocity discrepancy between sources, and the identity-confidence value to three learned scalar weights that sum to unity and are conditioned on the current kinematic context and association-confidence score. The fused precision matrix and fused state are then computed as

$$P_f^{-1} = \omega_{pred} P_{pred}^{-1} + \omega_r P_r^{-1} + \omega_i P_i^{-1} \tag{1}$$

$$P_f = (P_f^{-1} + \varepsilon I)^{-1} \tag{2}$$

$$x_f = (\omega_{pred} \hat{x}_{pred} + \omega_r x_r + \omega_i x_i) / (\omega_{pred} + \omega_r + \omega_i + \varepsilon) \tag{3}$$

Because P_f is a sum of positive semi-definite matrices weighted by non-negative scalars, it is itself positive semi-definite by construction, so fusion always reduces uncertainty relative to any single source, matching the consistency guarantee of classical Covariance Intersection while allowing the mixing weights to adapt to context. Four operational cases are handled explicitly: (i) both sensors fresh, giving a full three-way fusion; (ii) radar fresh with IFF held, giving a radar-weighted update; (iii) IFF fresh with radar held, giving an IFF-weighted update; (iv) both sensors held, giving a kinematic coast with frozen covariance to prevent divergence during extended sensor gaps. Handling these cases explicitly, rather than relying on the learned weights alone to suppress a stale source, avoids a failure mode in which a held measurement is inadvertently treated as a fresh, informative observation.

F. Track Management

The track manager implements a full track lifecycle using a bounded hit-score system. Each system track carries two bounded integer scores, one per sensor(radar an IFF), incremented on a consistent same-sensor-identity association and decremented on a miss. A track is confirmed once the combined score reaches a minimum threshold,

$$radar\_hit\_score + iff\_hit\_score \geq Amin = 3 \tag{4}$$

and confirmation is sticky, meaning that once a track is confirmed it is never reversed again until the track is deleted. This prevents a single missed scan from spuriously reverting a well-established track to tentative status. Deletion follows a differential, per-sensor logic in which miss counters are incremented only at the respective sensor's own scan times and only while the track holds an active sensor identity. For a joint track, disassociation proceeds in stages: if the consecutive-miss count for one sensor exceeds a maximum threshold, that sensor's identifier is cleared and the track degrades to single-sensor mode without deleting the underlying system track, while remaining eligible to re-accept that sensor if a new match subsequently arrives; full deletion occurs only once both sensors have been disassociated simultaneously. For a track that already holds only a single sensor identity, deletion follows directly once that sensor's consecutive-miss count exceeds the threshold. Track births occur for unmatched radar and IFF detections, with a newly born tentative track beginning at the minimum hit score and requiring the same combined threshold to reach confirmation. The resulting confirmed track set, together with its fused state, covariance, identity history, and sensor-freshness flags, is fed back into the following scan's prediction step, closing the loop shown in Fig. 1.

III. NEURAL NETWORK AND TRAINING METHODOLOGY

A. Training Scenario

Figure 2 shows the three-phase training pipeline, beginning with association-module training and proceeding through two fusion-module training phases before the final joint fine-tuning phase. This staged curriculum was adopted because jointly training all modules from a random initialization proved unstable in preliminary experiments. The association modules and the fusion module each depend on the other producing reasonable outputs, so training them simultaneously from scratch tends to converge to a degenerate solution in which one module compensates for the other's errors rather than both learning their intended function. Staging the curriculum, and freezing modules that are not the current training target, breaks this circular dependency.

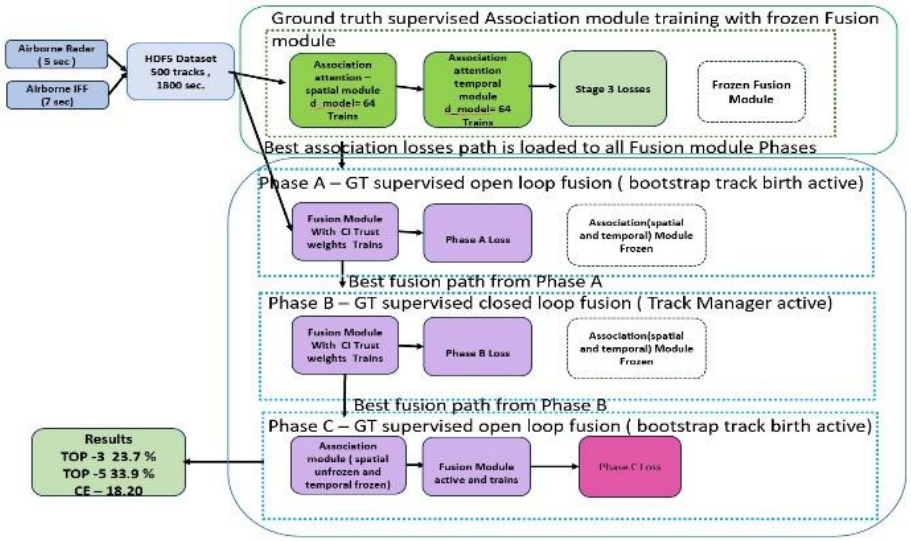
Ground truth supervised Association module training with frozen Fusion module
Airborne Radar (5 sec)
Airborne IFF (7 sec)
HDF5 Dataset 500 tracks , 1800 sec.
Association attention – spatial module d_model= 64 Trains
Association attention temporal module d_model= 64 Trains
Stage 3 Losses
Frozen Fusion Module
Best association losses path is loaded to all Fusion module Phases
Phase A – GT supervised open loop fusion (bootstrap track birth active)
Fusion Module With CI Trust weights Trains
Phase A Loss
Association(spatial and temporal) Module Frozen
Best fusion path from Phase A
Phase B – GT supervised closed loop fusion (Track Manager active)
Fusion Module With CI Trust weights Trains
Phase B Loss
Association(spatial and temporal) Module Frozen
Best fusion path from Phase B
Phase C – GT supervised open loop fusion (bootstrap track birth active)
Association module (spatial unfrozen and temporal frozen)
Fusion Module active and trains
Phase C Loss
Results
TOP -3 23.7 %
TOP -5 33.9 %
CE – 18.20

Fig. 2 Three-phase training pipeline: association-module training, followed by open-loop and closed-loop fusion pre-training, followed by joint end-to-end fine-tuning.

B. Association Module Training

Both the spatial and temporal association modules are jointly trained with the fusion module frozen, using ground-truth tracks for supervision. The composite loss combines a primary cross-entropy association term with a gated ground-truth supervision term, dense binary cross-entropy terms computed over hard-negative candidate pairs, an identity-consistency term, null-score and miss penalties, an entropy term that encourages peaked association distributions, spatial- and joint-plausibility terms, and a set of parameter-level regularizers that anchor the learnable null score, preventing pointer-head bias collapse, discouraging excessive null-token usage, and preventing identity-temporal weight collapse during radar-only timesteps. Each term addresses a specific failure mode observed during development. For example, the miss penalty specifically targets the case in which a valid, in-gate candidate exists but the network nonetheless assigns the highest score to the null token, which the cross-entropy term alone does not sufficiently discourage once the network has learned to be conservative.

C. Open-Loop Fusion Pre-Training

During this phase of training the fusion module is trained in isolation, with both association modules frozen and initialized from the association-training checkpoint. Ground-truth position labels supervise the three-way fusion through a weighted combination of a clamped mean-squared position-error term, a normalized estimation-error-squared consistency term that penalizes deviation from its chi-squared expected value, and a reconstruction term that discourages the fused state from drifting far from both sensor measurements simultaneously. Training uses the

AdamW optimizer, selected because it decouples weight decay from the gradient update and therefore regularizes the trust-head weights without distorting the effective learning rate, together with a step-decay learning-rate schedule.

D. Closed-Loop Fusion Pre-Training

The track manager is activated in this phase, so track births, confirmations, disassociations, and deletions occur during training exactly as they would at inference time, while the association modules remain frozen so that only the fusion module is updated. The loss now shifts to a closed-loop and is composed of ground-truth-supervised position-accuracy term as the primary signal, complemented by the reconstruction term, a smoothness term that penalizes frame-to-frame discontinuities in the fused trajectory, and the consistency term from the previous phase at reduced weight, since positional accuracy is treated as the primary objective at this stage and covariance consistency as secondary. Training this phase in closed loop, rather than continuing to assume ground-truth track continuity, exposes the fusion module to the same track-fragmentation and re-birth events that it will encounter at test time, which open-loop training cannot imitate.

E. Joint End-to-End Fine-Tuning

In the final phase, all three modules are trained jointly, without ground-truth association labels, in a fully closed loop. The composite loss combines a self-supervised cross-entropy term, in which spatial proximity to the nearest radar detection serves as a surrogate label so that no explicit association annotation is required, together with the reconstruction and smoothness terms carried over from the previous phase, an uncertainty term, an entropy term encouraging confident association decisions, and a joint-assignment term that rewards simultaneous radar and IFF assignment to the same system track. The reconstruction term provides the only gradient path from the association output into the fusion module's trust-head weights without ground-truth positions, flowing from the fused state through the covariance-intersection computation and the trust head back to the association logits. The learning rate is halved midway through this phase once the joint training has largely converged, allowing the network to settle into a stable local optimum without oscillating.

IV. EXPERIMENTAL EVALUATION AND RESULTS

A. Simulation Scenario Design and Dataset

Existing open-source benchmark datasets lack asynchronous radar-IFF aerial target tracks with heterogeneous sensor-observation distributions spanning joint, radar-only, and IFF-only measurements. Therefore, generating a synthetic dataset was a genuine requirement, so accordingly a dataset was constructed to replicate the operational

environment of an airborne surveillance system tracking multiple, dense, highly dynamic targets. The track-generation module simulates fighter-class aircraft using a six-degree-of-freedom aerodynamic model with stochastic maneuvering, drawing linear acceleration and clipped subsonic speed from ranges representative of fighter and transport aircraft, with initial positions distributed across a wide-area, high-altitude operational airspace. Three track cohorts simulate realistic engagement dynamics. A majority cohort that persists throughout the scenario, a cohort that terminates partway through as targets exit sensor coverage, and a cohort that enters partway through as new targets are detected. This structure produces crossing tracks among simultaneous targets with random initial bearings, and the combination of late entries, early departures, and stochastic maneuvering both creates association ambiguity and violates the fixed-model assumptions on which classical Kalman-filter baselines rely. Sensor availability is predominantly joint radar-and-IFF, with smaller fractions of radar-only coverage, representing non-cooperative aircraft, and IFF-only coverage, representing radar-masked aircraft. Detection probabilities, clutter rates, a bounded miss tolerance, burst dropouts, sensor degradations, occasional wrong-identity events, and periodic identity-code refreshes together simulate the contested, electronic-warfare-affected conditions that motivate the fusion problem in the first place, rather than the comparatively benign, low-clutter conditions under which classical association methods are typically validated.

B. Complementary Sensor Noise Model

A key design principle of the simulation, and the primary motivation for multi-sensor fusion in this setting, is the deliberately complementary noise structure of the two sensors, implemented through a Jacobian-propagated covariance model that encodes sensor-specific measurement anisotropy grounded in physical measurement principles. Radar provides higher azimuth accuracy but degraded elevation accuracy, consistent with a mechanically or electronically scanned two-dimensional radar whose elevation estimate is derived rather than directly measured, and also provides three-dimensional velocity through Doppler processing, with vertical velocity substantially more uncertain than the horizontal components. IFF provides superior elevation accuracy but inferior azimuth accuracy, creating the complementary coverage that motivates fusion. When the identity code is stable, the IFF elevation variance is tightly clamped, and a dedicated barometric-altitude channel, absent from radar, provides an additional elevation-accurate observation. When the identity code is unstable, the elevation variance is instead relaxed substantially to reflect the reduced confidence in that scan. IFF carries no Doppler velocity information, so its velocity

uncertainty is set far higher than that of radar's, ensuring that the fusion module learns to defer to radar for all velocity components. Table 1 summarizes this complementary structure and characterizes the dataset.

Table 1 Complementary sensor noise model (asymmetric structure motivates fusion).

Parameter	Radar	IFF	Fusion benefit
Range / azimuth noise	Lower azimuth error	Higher azimuth error	Radar preferred (lateral)
Elevation noise	Poor (derived estimate)	Good (identity-stable)	IFF preferred (vertical)
Altitude (direct)	None	Direct barometric channel	IFF unique elevation channel
Velocity noise	Low (Doppler-derived)	High (no Doppler)	Radar preferred (all velocity)
Identity code	None	Transponder identity code	IFF unique identity signal
Scan rate / detection probability	Faster, higher	Slower, lower	Asynchronous; both needed

C. Test Scenario and Metrics

The held-out test scenario comprises one hundred airborne targets over a twenty-minute engagement window. The majority of targets are present throughout and smaller cohorts terminating early or appearing late. High percentage of scans contain joint radar-and-IFF tracks to validate fusion opportunities alongside a small number of radar-only and IFF-only scans. Model weights were loaded from the best joint-fine-tuning checkpoints and frozen for evaluation. Five per-scan metrics are analyzed at a fixed association gate i.e target detection rate, false-track rate, gated root-mean-square error, the Optimal Sub-Pattern Assignment metric [5, 10], and the joint-fusion rate, providing a view of the framework's performance that spans detection completeness, false-track generation, positional accuracy, and overall multi-object tracking quality.

D. Final Track State

At the end of the scenario the majority of alive system tracks are confirmed, and the great majority of confirmed tracks are jointly fused from both sensors. The ratio of system tracks to ground-truth targets substantially exceeds one, reflecting multiple track segments per ground-truth target due to track re-birth following sensor gaps rather than a proliferation of spurious tracks, since the false-track rate remains low and the great majority of confirmed tracks correspond to real targets. Figure 3 shows the track count growing over the course of the scenario without runaway growth, with a visible step where late-appearing targets enter the scenario.

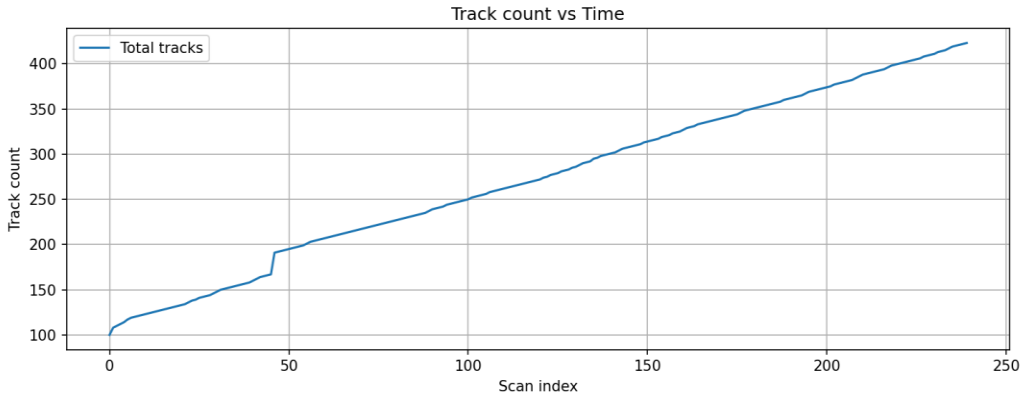

Fig. 3 Track count versus scan number, showing growth from the initial target population without unbounded track proliferation.

E. Baseline Comparison

Three classical baselines are evaluated on the identical held-out dataset: a Global-Nearest-Neighbor association method combined with an extended Kalman filter and simple-average IFF fusion; an Unscented Kalman Filter combined with Global-Nearest-Neighbor association; and a fixed-weight Covariance Intersection baseline that pairs Global-Nearest-Neighbor association with the classical, non-adaptive fusion weight, serving as a direct ablation of the learned fusion module's contribution. A joint probabilistic data association baseline was excluded after it produced a near-total loss of track continuity under the high-clutter conditions of the test scenario, which is itself informative: it indicates that the scenario is deliberately constructed to be more demanding than the conditions under which such classical association methods are typically evaluated.

Table 2 Baseline comparison summary (identical held-out test scenario; boldface indicates the best value per column).

Method	**Detection rate(TDR&)**	**False-track rate(FTR %)**	**RMSE(m)**	**OSPA**
Global-Nearest-Neighbor + EKF	99.2	0.7	41.94	87.12
Unscented Kalman Filter + GNN	99.2	2.3	39.71	87.72
Fixed-weight Covariance Intersection	99.2	3.5	38.83	88.77
Proposed framework (learned fusion)	**100.0**	**0.7**	**51.56**	**162.51**

F. Results and Discussion

The proposed framework achieves the highest detection rate among all methods compared, with zero missed targets across the full evaluation window, while all baselines miss a small but consistent fraction of detections per scan. In a dense, high-value airborne surveillance scenario a missed target generally cannot be recovered within the

same engagement cycle, so this margin is operationally significant even where the absolute gap in detection rate appears numerically modest. The false-track rate of the proposed framework ties the best-performing baseline, while the remaining two baselines generate substantially more false tracks, reflecting weaker association discrimination under dense, ambiguous conditions.

The gated root-mean-square error of the proposed framework is higher than the classical baselines, and three factors plausibly account for this gap. First, the framework maintains multiple track segments per ground-truth target as a consequence of its deliberate track-history-preservation design, so newly reborn segments begin without a filter warm-up period, which temporarily elevates instantaneous position error immediately following a re-birth event. Second, the fusion module's covariance estimate is overconfident on held-out data relative to its chi-squared expected value, which slows the effective incorporation of new measurements relative to a filter with well-calibrated covariance. Third, the classical baselines use hand-calibrated, fixed measurement-noise parameters tuned specifically to the evaluation scenario, whereas the covariances used by the proposed framework are entirely learned from training data and have not been separately recalibrated for the test distribution. A per-sensor breakdown of the neural-network root-mean-square error, shown in Fig. 4, indicates that radar-only tracking is more accurate than IFF-only tracking, and that the fused estimate falls between the two, consistent with the complementary but unequal per-sensor accuracy summarized in Table 1.

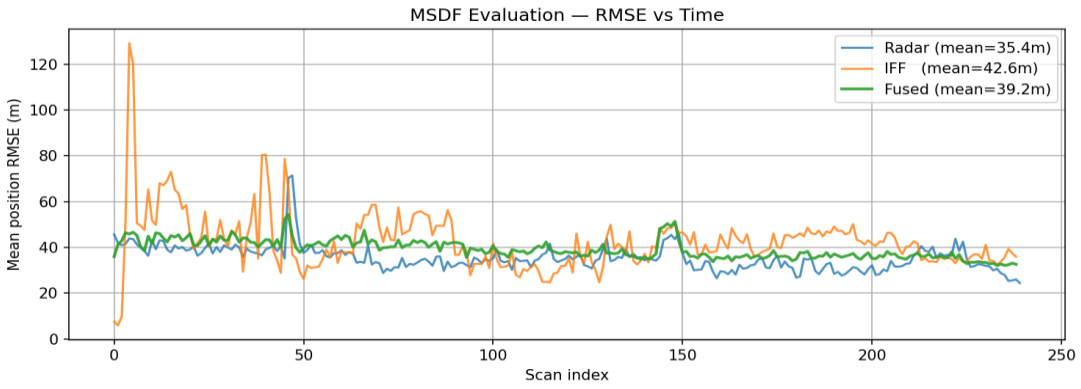

Fig. 4 Neural-network root-mean-square error versus scan number for radar-only, IFF-only, and fused state estimates.

G. OSPA Interpretation

The Optimal Sub-Pattern Assignment metric for the proposed framework is substantially higher than for the baselines, dominated by a cardinality mismatch between the larger number of system tracks maintained and the smaller number of ground-truth targets alive at any given scan. Because the false-track rate remains low, confirming that the

great majority of system tracks correspond to real targets, the excess cardinality is attributable to track fragmentation rather than false alarms. The metric penalizes multiple track segments representing the same ground-truth target as if they were independent extraneous tracks, even though each segment is individually a correct, low-error representation of that target during the interval it covers. This reflects an inherent trade-off between the history-preserving design adopted here, which favors higher detection reliability and target-identity continuity, and a more aggressive pruning strategy that would reduce the Optimal Sub-Pattern Assignment score at the cost of a higher missed-target rate. It suggests that cardinality-aware or track-continuity-aware variants of the metric may be more informative than the standard formulation for evaluating frameworks with this design philosophy.

H. Filter Consistency and Computational Speed

The normalized estimation-error-squared values for radar-only, IFF-only, and fused tracks are all substantially above the chi-squared expected value corresponding to the state dimension, indicating that Phase C training, shown in Fig. 5, optimized primarily for association accuracy rather than for covariance calibration. This overconfidence does not directly affect the detection-rate or false-track-rate metrics, which depend on the association decisions rather than on the covariance magnitude. However, it does affect any downstream process that consumes the reported covariance as a measure of confidence, such as a threat-prioritization or resource-allocation stage that would otherwise treat an overconfident covariance as more informative than it actually is. Consistency regularization of the fusion covariance is accordingly identified as a primary direction for future work. On computational speed, the framework achieves a modest average frame rate on central-processing-unit-only inference in a Python implementation, which is adequate for offline evaluation. However, with the incorporation of graphics-processing-unit deployment for real-time operation based on comparable transformer-based multi-object trackers reported in the literature [11], an order-of-magnitude speedup is very much possible without architectural changes, though this has not yet been measured directly for the present framework and should be validated before any real-time deployment claim is made.

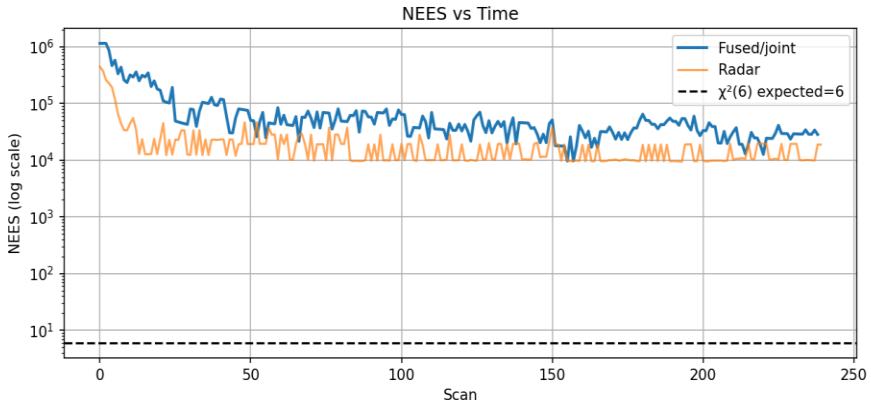

Fig. 5 Normalized estimation-error-squared versus scan number (log scale), compared against the chi-squared expected value.

I. Limitations

Three limitations of the present study are worth stating explicitly. First, the evaluation is conducted on a single synthetic held-out scenario with a fixed random seed. While the scenario is designed to be dense and operationally representative, repeated evaluation across multiple random seeds and across scenarios with different target densities and maneuver profiles would give a more robust estimate of the framework's performance distribution than a single scenario can provide. Second, the covariance-calibration gap documented in Section IV.H means that the framework's reported uncertainty should not yet be trusted as an absolute confidence measure by any downstream consumer, even though the point estimates and association decisions it produces are reliable. Third, the framework has been evaluated exclusively in simulation against a two-sensor radar-IFF configuration. Validation against recorded or field sensor data, and extension to additional sensor types, remain necessary steps before the framework could be considered for operational deployment.

V. CONCLUSION

This paper presented an end-to-end transformer-based multi-sensor data fusion framework for an airborne surveillance system, combining spatial and temporal cross-modal association with adaptive, learned Covariance Intersection fusion and a principled track lifecycle manager. The multi-term association loss provides comprehensive ground-truth-supervised training for the association stage, staged open-loop and closed-loop pre-training establishes a well-behaved starting point for the fusion module, and self-supervised joint fine-tuning subsequently refines all modules together without requiring ground-truth association labels. On a held-out, densely populated test scenario, the proposed framework achieves the highest detection rate among all methods compared, including three classical baselines, while maintaining a false-track rate comparable to the best classical alternative. These results indicate that

learned, adaptive fusion weights and joint, transformer-based association offer a meaningful reliability advantage over fixed-model and fixed-weight classical approaches in dense, asynchronous, multi-sensor tracking scenarios, at the cost of a currently uncalibrated covariance estimate and increased track fragmentation. All these learnings, motivate the future work which has been truly outlined above, namely covariance-consistency regularization, extension to an event-driven framework supporting multiple radars and additional cooperative sensors such as Automatic Dependent Surveillance-Broadcast(ADSB) and Automatic Identification System(AIS) feeds, and validation on recorded field data.

References

[1] Bar-Shalom, Y., and Tse, E., "Tracking in a Cluttered Environment with Probabilistic Data Association," Automatica, Vol. 11, No. 5, 1975, pp. 451–460.

[2] Reid, D., "An Algorithm for Tracking Multiple Targets," IEEE Transactions on Automatic Control, Vol. 24, No. 6, 1979, pp. 843–854.

[3] Brasó, G., and Leal-Taixé, L., "Learning a Neural Solver for Multiple Object Tracking," Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020, pp. 6247–6257.

[4] Blackman, S., and Popoli, R., Design and Analysis of Modern Tracking Systems, Artech House, Norwood, MA, 1999.

[5] Schuhmacher, D., Vo, B.-T., and Vo, B.-N., "A Consistent Metric for Performance Evaluation of Multi-Object Filters," IEEE Transactions on Signal Processing, Vol. 56, No. 8, 2008, pp. 3447–3457.

[6] Bar-Shalom, Y., Li, X. R., and Kirubarajan, T., Estimation with Applications to Tracking and Navigation, Wiley, New York, 2001.

[7] Julier, S. J., and Uhlmann, J. K., "A Non-Divergent Estimation Algorithm in the Presence of Unknown Correlations," Proceedings of the American Control Conference, 1997, pp. 2369–2373.

[8] Vaswani, A., Shazeer, N., Parmar, N., et al., "Attention Is All You Need," Advances in Neural Information Processing Systems, Vol. 30, 2017.

[9] Verma, R., Rajesh, R., and Easwaran, M. S., "Modular Multi-Target Tracking Using LSTM Networks," arXiv:2011.09839, 2020.

[10] Rahmathullah, A. S., García-Fernández, A. F., and Svensson, L., "Generalized Optimal Sub-Pattern Assignment Metric," Proceedings of the International Conference on Information Fusion, 2017, pp. 1–8.

[11] Meinhardt, T., Kirillov, A., Leal-Taixé, L., and Feichtenhofer, C., "TrackFormer: Multi-Object Tracking with Transformers," Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022, pp. 8844–8854.

[12] Zeng, F., Dong, B., Zhang, Y., et al., "MOTR: End-to-End Multiple-Object Tracking with Transformer," Proceedings of the European Conference on Computer Vision, 2022.

[13] Pinto, J., Hess, G., Ljungbergh, W., et al., "Deep Learning for Model-Based Multi-Object Tracking," IEEE Transactions on Aerospace and Electronic Systems, 2023.

[14] Wei, X., et al., "Transformer-Based Multi-Sensor Hybrid Fusion for Multi-Target Tracking," Proceedings of the IEEE International Conference on Information Fusion, 2024.

[15] Blom, H. A. P., and Bar-Shalom, Y., "The Interacting Multiple Model Algorithm for Systems with Markovian Switching Coefficients," IEEE Transactions on Automatic Control, Vol. 33, No. 8, 1988, pp. 780–783.

[16] Blasch, E., et al., "Machine Learning/Artificial Intelligence for Sensor Data Fusion," IEEE Aerospace and Electronic Systems Magazine, Vol. 36, No. 7, 2021, pp. 80–93.

[17] Sharma, G., Kumar, C. R. S., and Rajesh, R., "Multi-Sensor Data Fusion for Aerial Target Tracking Using Transformer-Based Association Attention and Self-Supervised Fine-Tuning," Proceedings of the IEEE MAPCON Conference, Nagpur, India, Dec. 2026.